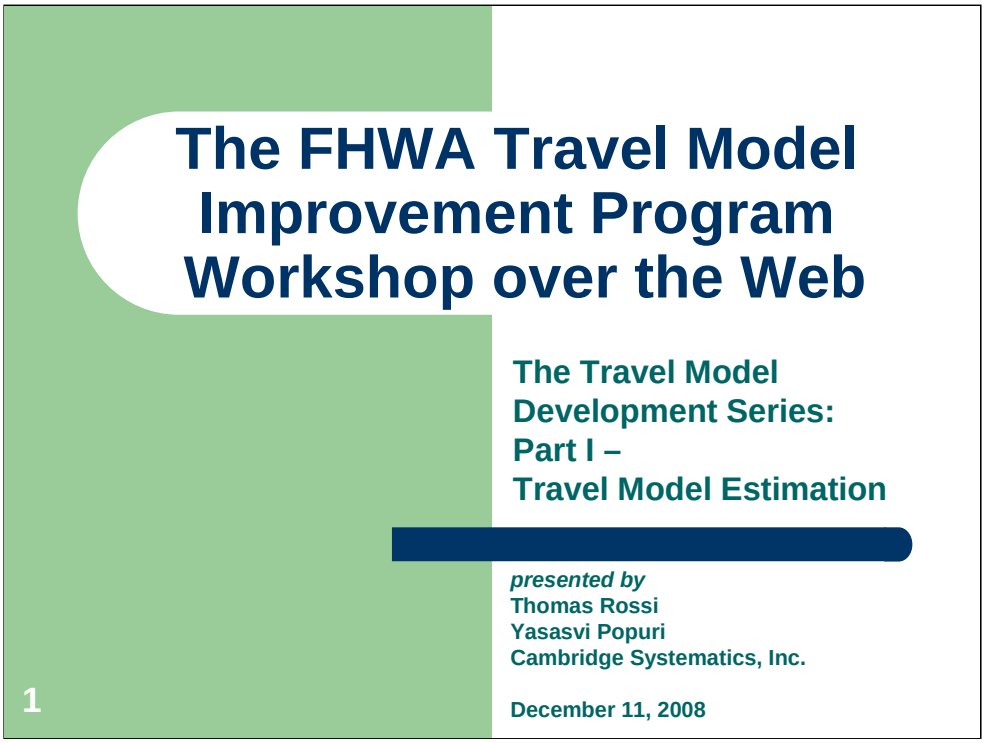
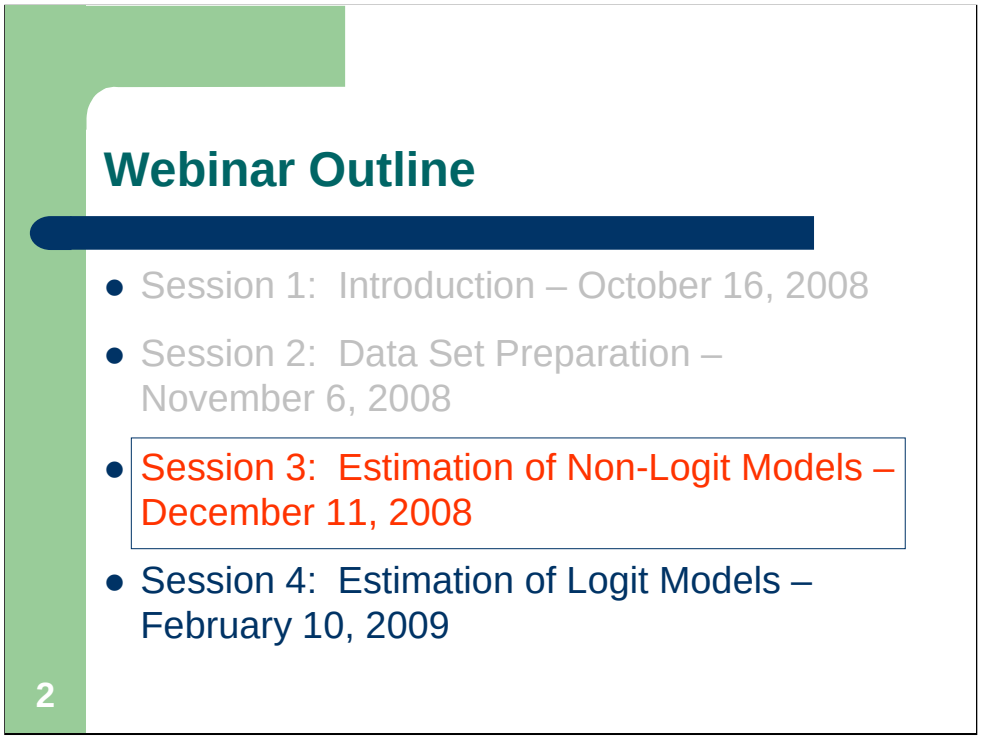


Key Message: Purpose of the Webinar Series

Details:

Welcome to the FHWA TMIP Workshop over the Web. This workshop is targeted at Transportation modelers who have a low to moderate level of familiarity with the estimation and validation of travel models.

This series of webinars will introduce the development of model estimation data sets, the structures of the various model components, and the procedures for estimating models. The workshop will include lectures, discussion, and "homework," that participants will be expected to complete between sessions.

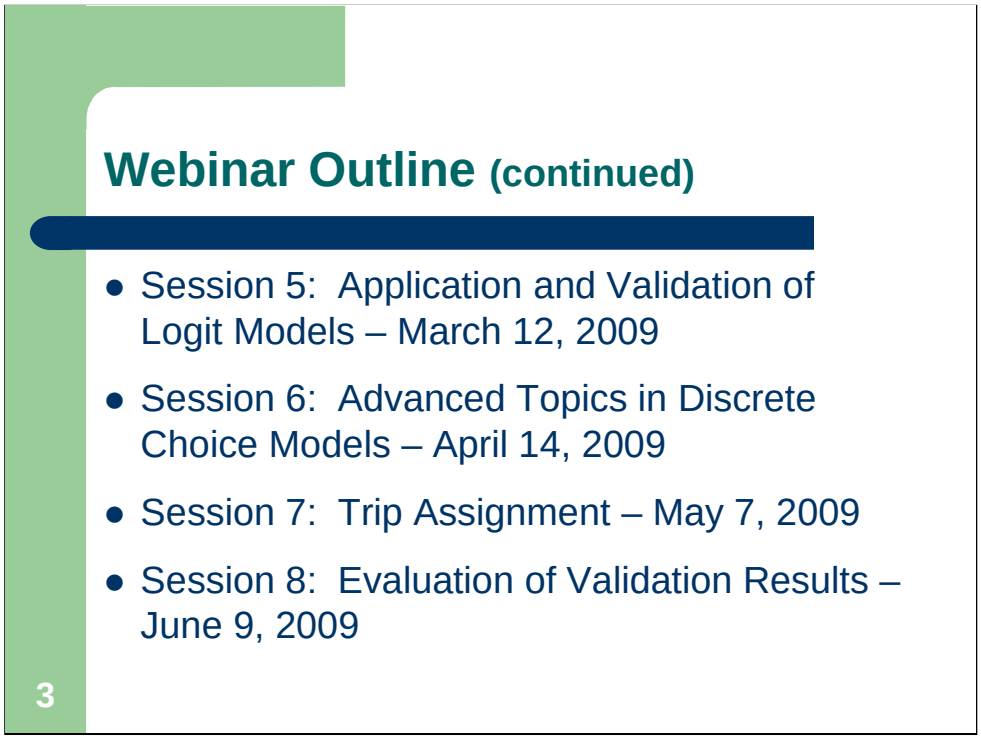


Key Message: Current Session

Details:

This session deals with the estimation of non-logit models, the typical model types and data sources.

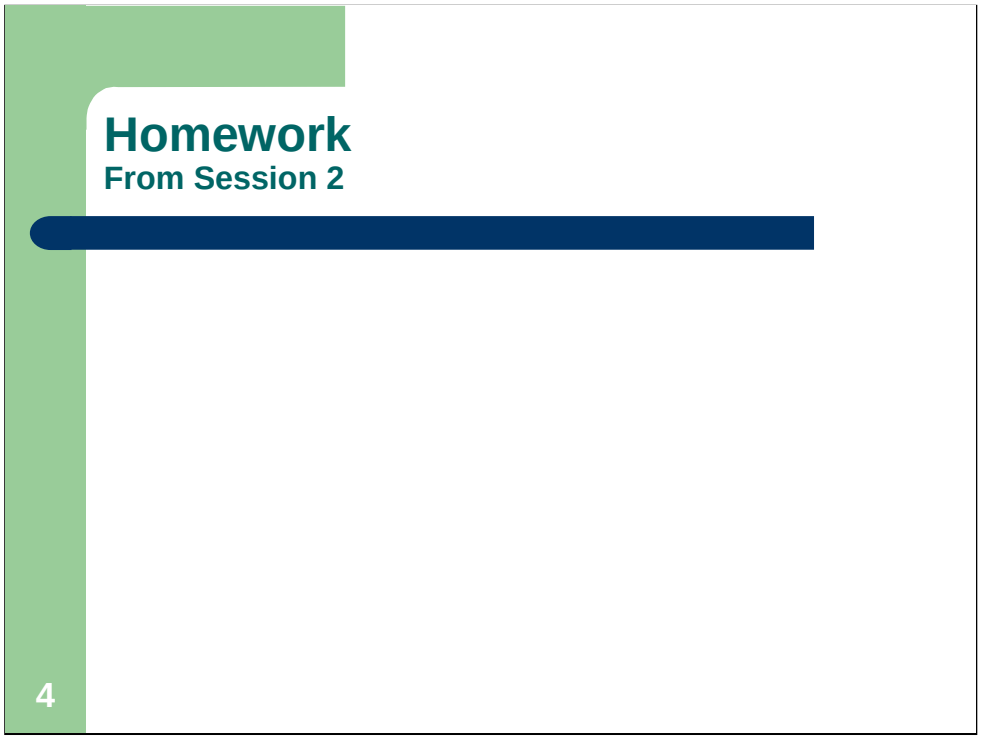
Session 4, which will be conducted on February 10, 2009 will cover the various aspects of logit model estimation.



Key Message: Upcoming Sessions

Details:

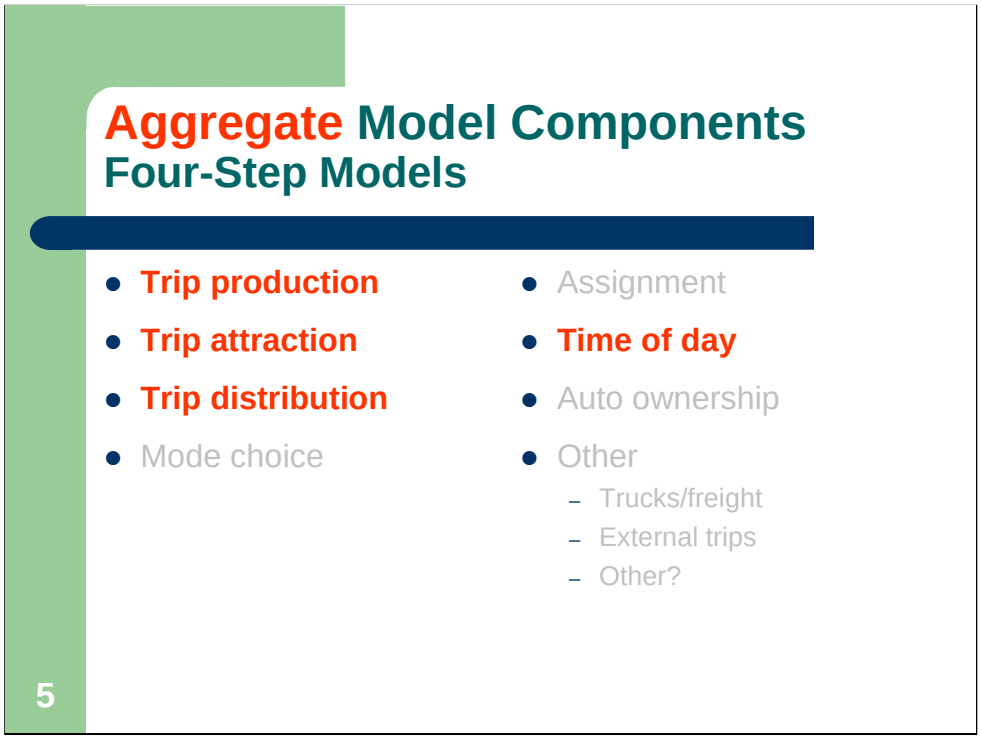
The dates for Sessions 5-8 have been determined. Session 5 will be conducted on March 12, 2009; Session 6 will be conducted on April 14, Session 7 on May 7, and Session 8 on June 9.



Key Message: Homework 3 Discussion

Details:

Please refer to the homework solutions posted at the website.

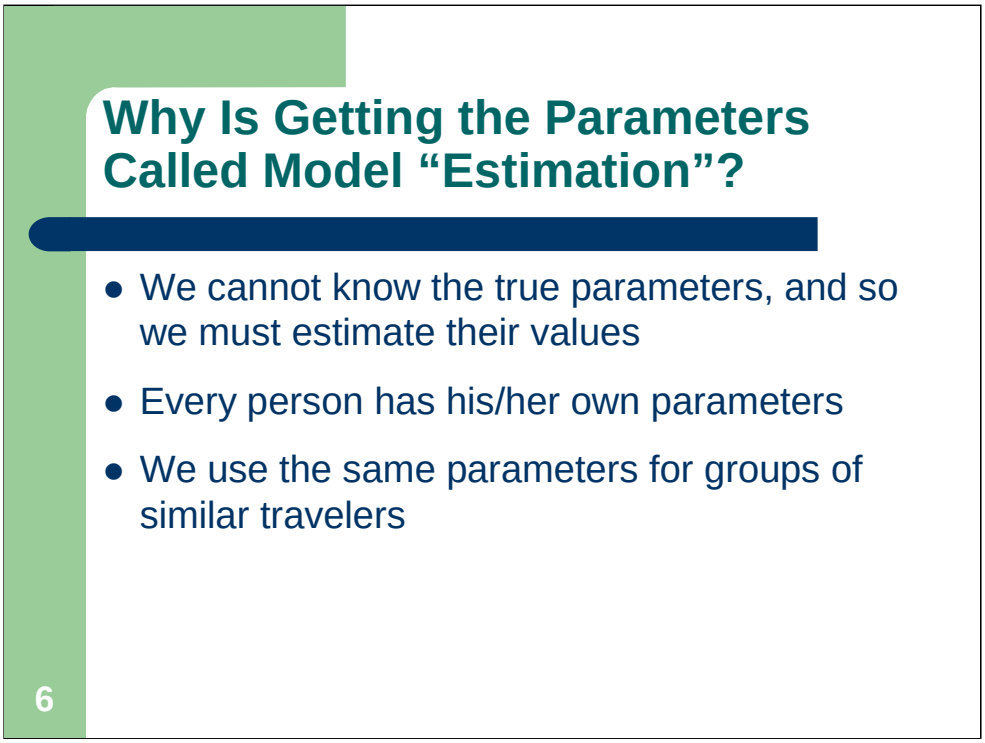


Key Message: Aggregate Model Components

Details:

In this session we will deal with aggregate components of travel demand models. By "aggregate" we refer to models that are estimated for groups of trip records lumped together, as opposed to disaggregate models such as logit models where each record is considered individually in model estimation. Disaggregate models will be covered in Session 4.

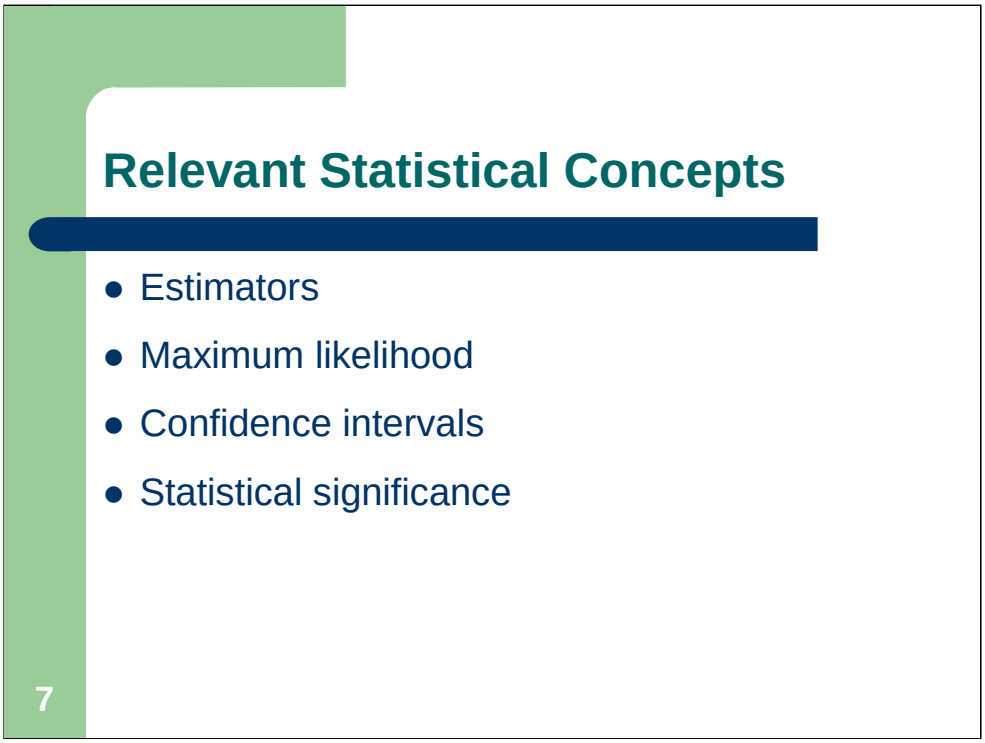
Aggregate models include the trip generation models (production and attraction), the gravity model for trip distribution, and time of day factoring models. While there are disaggregate versions of all of these models, the aggregate versions are much mode common.



Key Message: Why is getting parameters called estimation?

Details:

Why is getting the parameters called model "estimation"? The fact is that we can never know the "true" values of model parameters; we can only observe the behavior (relative to a specific model component) of a sample of travelers, all of whom do not behave the same way. In an aggregate model, we assume that the same parameters apply to all members of groups of similar travelers. We use statistical methods to come up with our best estimates of the values of these "average" parameters.



Key Message: Statistical Concepts

Details:

While this webinar cannot teach a complete statistics course, there are some important statistical concepts that are relevant to the discussion.

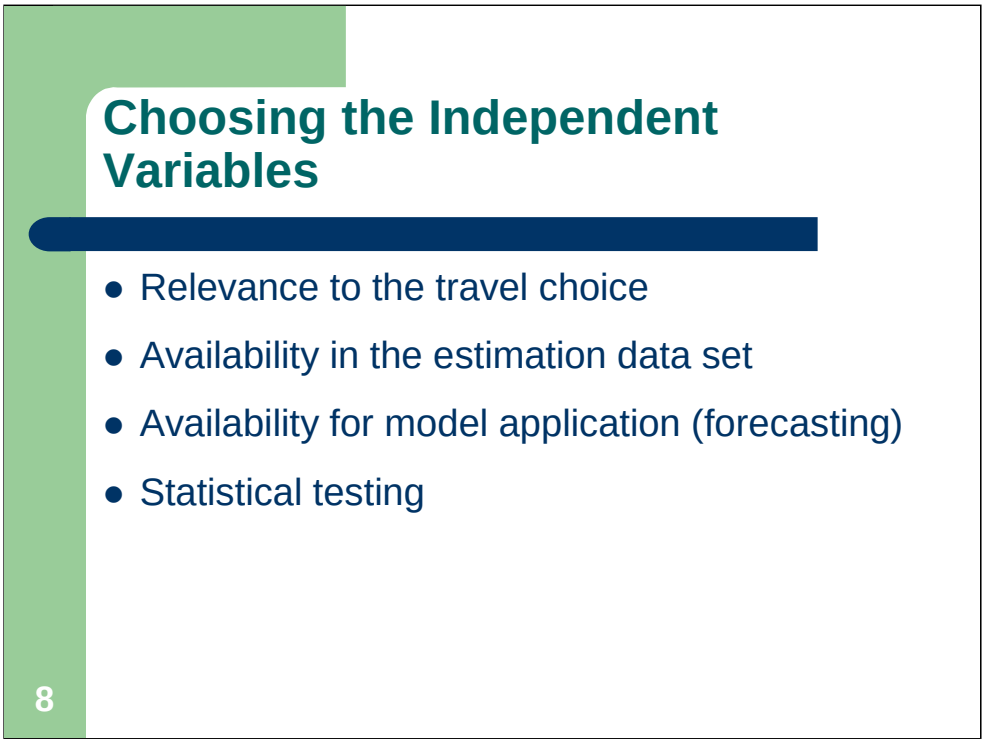
• An **Estimator** is the form of an expression for estimating an unknown parameter. For example, if there are n observations of a parameter, the sample mean, $1/n \Sigma X_i$, is an estimator.

• **Maximum likelihood** estimators are those that best (or are "most likely" to) explain the observed data. A likelihood function can be computed for a probability density function, and the maximum likelihood estimator for a parameter is that for which the likelihood function is maximized.

• A **confidence interval** is associated with a specific probability level. For example, a 95% confidence interval can be computed for a parameter estimate; there is a 95% probability that the true parameter value lies within the confidence interval.

• The **level of significance** for our purposes refers to the probability that the value of the parameter is nonzero. For example, if an estimate is significant at the 5% level, this means that there is a 95% probability that the true parameter value is nonzero. Computing the level of significance requires statistical hypothesis testing.

Generally, the software used to estimate models computes the values described on this slide.



Key Message: Independent Variables

Details:

The independent variables are those on which the travel choice is based. For example, home based non-work trip attractions may be based on retail employment, non-retail employment, and households.

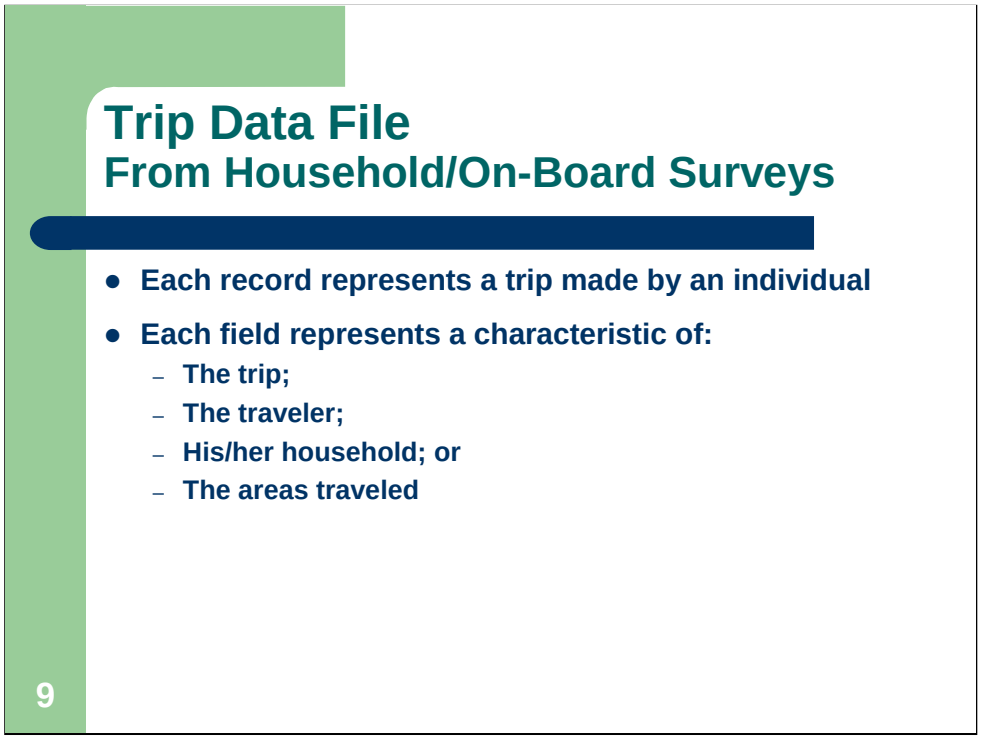
The important factors in choosing the independent variables include:

• The relevance of the variables in explaining the travel choice being modeled.

• The availability of the variables in the data set. For example, if income was not asked in a survey, it is not available to be used in models estimated from that survey data set.

• The availability of the variable for forecasting. For example, if income is asked in the survey but income forecasts are not available, it cannot be used.

• The results of statistical tests that demonstrate that the variable helps explain the travel choice.

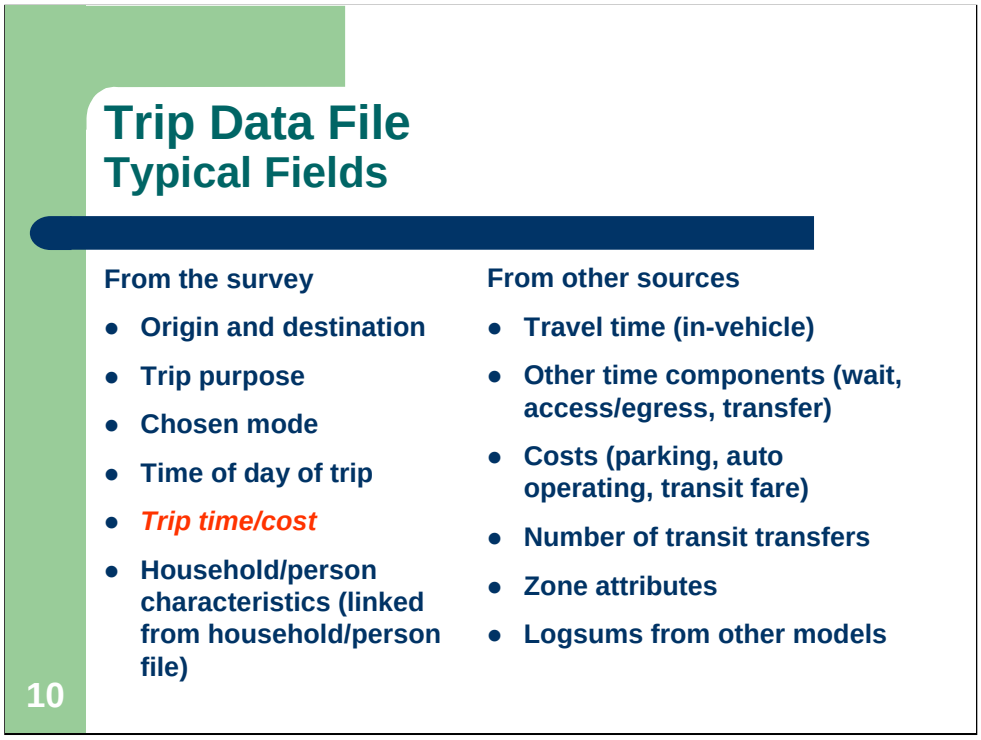


Key Message: Trip Data File

Details:

Where do you get data for model estimation?

The models we will discuss today use data from the trip files from surveys. Last session we discussed creating these trip files. You may recall that each record represents a trip, and each field represents an attribute of the trip, traveler, household, or area.



Key Message: Trip Data File Contents

Details:

Some of the typical fields in a trip file are as follows:

1. Latitude and longitude information for the origin and destination of the trip. This information can be post-processed to attach TAZ information to the origin and destination.
2. Origin and destination activities, which can be used to determine the purpose of the trip itself.
3. The transportation mode chosen to make the trip.
4. The starting and ending times of each trip, which can be used to deduce the total reported time for each trip.
5. The household and person characteristics for the tripmaker.
6. In-vehicle and out-of-vehicle times can be attached to each trip record based on the origin and destination information generated in step 1. The sources of the in-vehicle and out-of-vehicle times are the highway and transit networks.
7. Zonal land use data for the origin and destination zones can also be attached to the trip files.
8. Finally, for model systems that use mode choice logsums, the logsums from the origin of the trip to each destination can also be attached to the trip record.

Should Weighted Data Be Used in Estimating Aggregate Models?

- Example:

Household	Trips	Sampled?	Weight
1	5	Yes	1.0
2	10	Yes	2.0
3	10	No	(not surveyed)

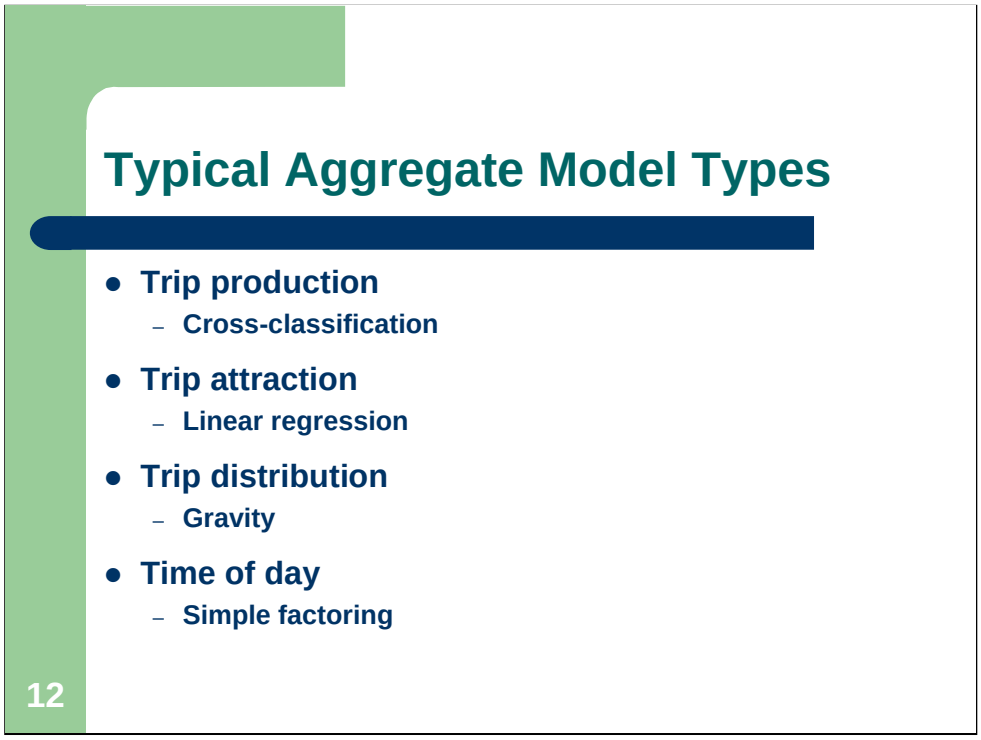
11

Key Message: Weighted Data

Details:

An important question for model estimation is whether data should be used in weighted or unweighted form. Recall that the survey data weights relate the survey observations to the population as a whole.

In general, it usually makes sense to use weighted data for aggregate model estimation. To illustrate why, consider the following simple example. Let's say that the entire population consists of three households, which produce trips at the levels shown on the slide. Further, assume that households 2 and 3 are in one market segment, and household 1 is in another. Now, say that the survey sample consists only of households 1 and 2. Therefore, the survey weight for household 2 is 2.0, since it represents 2 households while the weight for housheold 1 is 1.0. If we use unweighted data to estimate the average trip rate, we will get 7.5 while if we use weighted data, we will get 8.3 trips. It is easy to see that the rate obtained using the weighted data is a better estimate of the "true" trip rate.



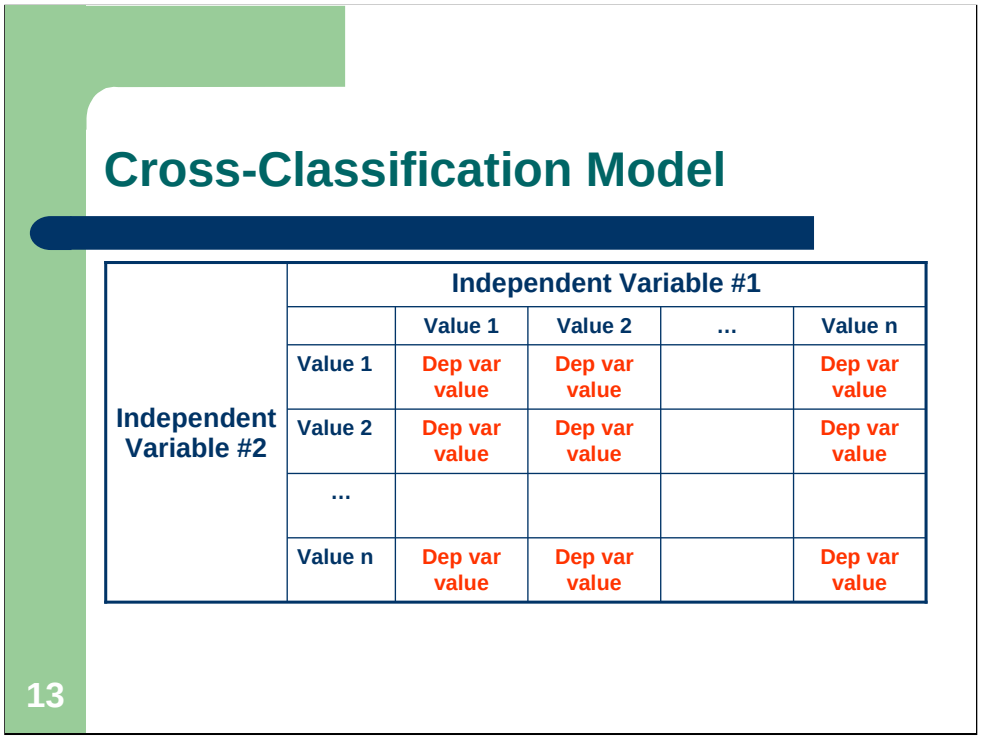
Key Message: Aggregate Models

Details:

Now we will discuss some specific model types.

These are the typical types of aggregate models used for each model component. Specifically, we will talk about four models:

1. Trip production, which uses cross-classification
2. Trip attraction, which uses linear regression
3. Trip distribution, which uses gravity model formulations
4. Time of day models, which use simple factoring



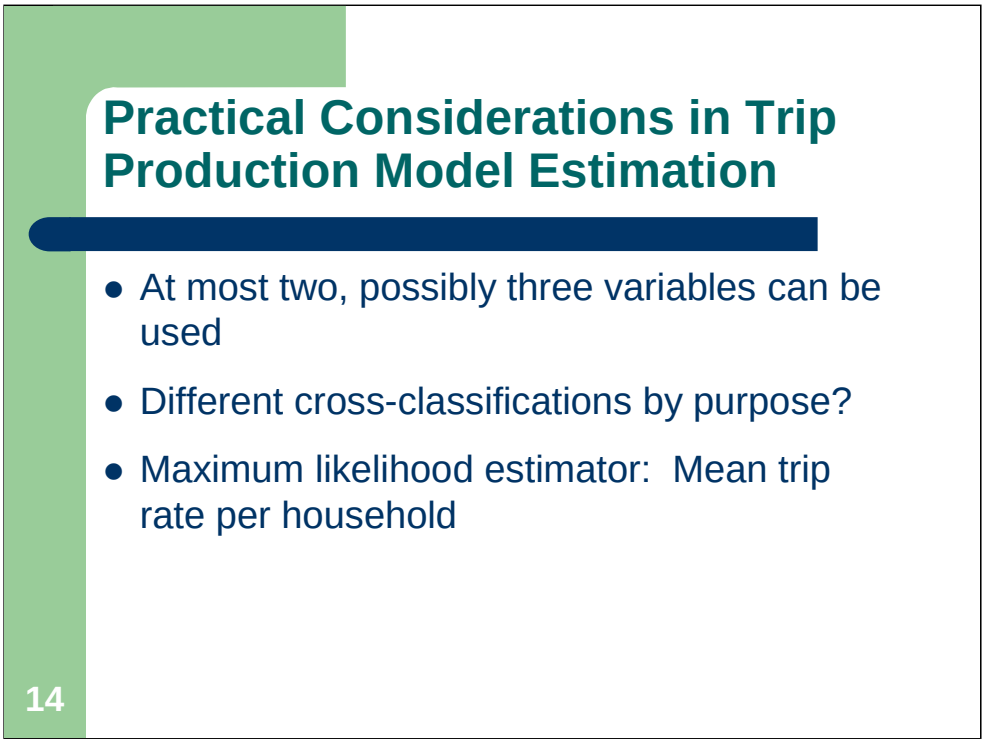
Key Contents: Cross-Classification Models

Details:

Cross-classification is the most commonly used technique for trip production models. Cross-classification involves the following items:

1. Identify independent variables that will be used to classify all households. Most common variables include household size and vehicles. The rationale is that households of a given size and with a given number of vehicles have similar necessity to make out-of-home trips. For example, all households with four people and two vehicles in the household are very similar in the number of trips they make each day. The higher the household size, the higher the necessity to make trips on an average and the higher the number of vehicles, the higher the number of vehicles, the higher the number of trips made. If there are 4 categories of household size and 4 categories of vehicle ownership, we have a total of 16 cross-class cells. That is, the matrix above will have 16 cells.
2. Identify the number of households that fall in each category. This is easily obtained from household surveys. For example, one can easily find how many households in the study region have four members in the household and 2 vehicles in the household. Similarly, the number of households that fall in each of the 16 cells discussed above can be obtained and tabulated.
3. Identify trips of each purpose made by the households of each type. That is, corresponding to each of the 16 cells above, one can find how many trips of each purpose are made. This can once again be obtained from household travel surveys.
4. Once we have the total trips of each type and the underlying number of households, we can find the trip rate for each purpose.
5. The outcome of this process is quite simply a set of production rate tables, one for each purpose. The cells in each table represent the average number of trips made by each household characterized by that cell.

The purpose of deriving these cross-class rates is that if one were able to forecast the number of households that fall in each category, one would also be able to forecast the number of trips that these households make by multiplying the total number of households with the trip rates.



Key Contents: Practical Considerations for Trip Production Models

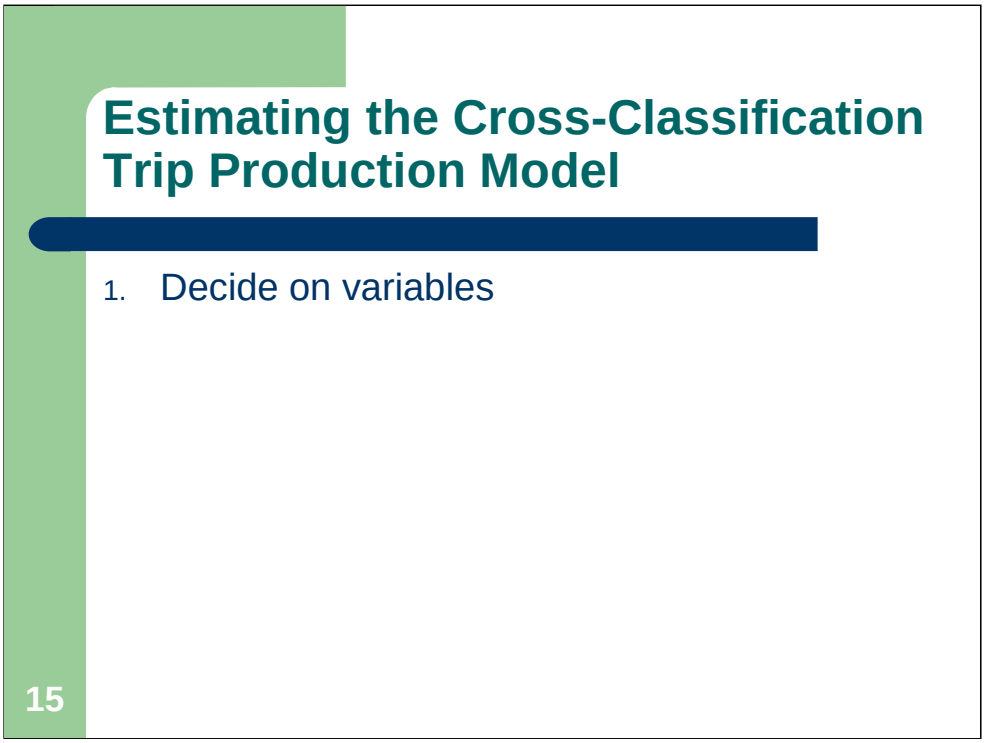
Details:

Some practical considerations for the cross-classification model. First, where do we get the data to estimate a cross-classification trip production model? But what kind of survey? The household survey, of course.

• With typical housheold survey sample sizes, there is seldom sufficient data to estimate trip rates for more than two, possibly three dimensions. For most trip purposes, where the number of total trips is relatively low, two dimensions are the most that can be used. So it is important to choose those variables wisely.

• Should the same variables be used for all trip purposes? This is up to those who estimate the models, those who will apply them, and those who will provide the data for application/forecasting. It is obvious that different variables are more important in explaining trip generation for different trip purposes. For example, the number of workers would likely be a key variable for home based work trips, but number of children would be more explanatory for home based school trips. However, the more different sets of variables used, the more different cross-classifications for which households will need to be forecast.

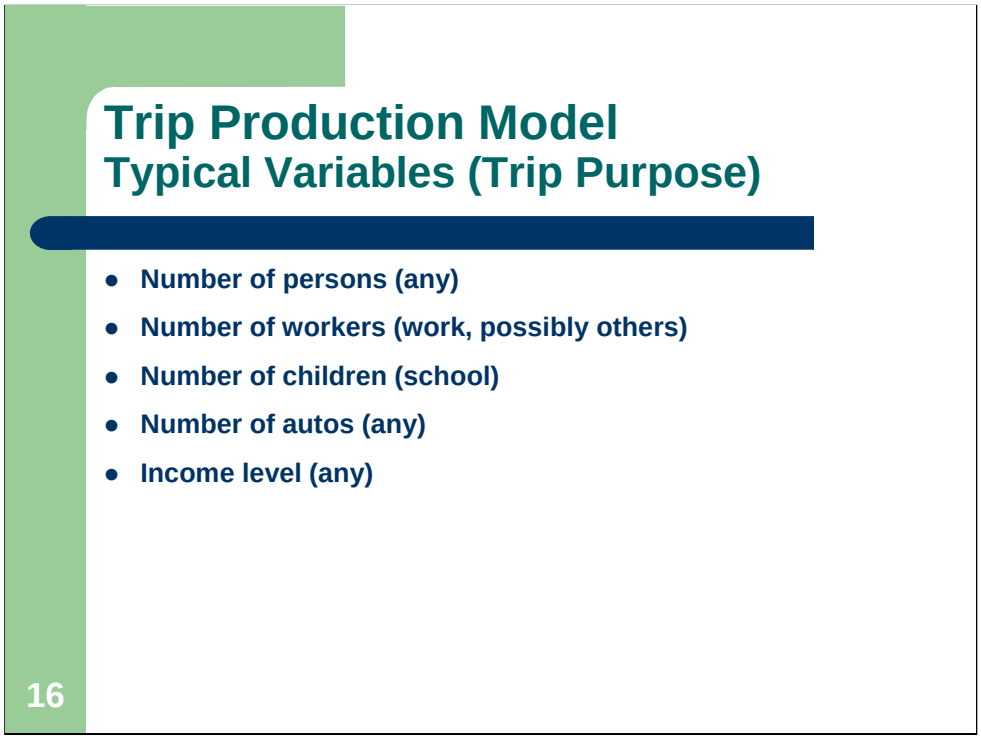
• The maximum likelihood estimator for trip generation is the mean trip rate for the population segment.



Key Contents: Variables to Use Trip Production Models

Details:

The first step in estimating the trip production model is to decide on the variables.



Key Contents: Variables to Use Trip Production Models

Details:

Some typical variables that can be considered are as follows:

•Household size, which can be used to estimate production rates for any purpose

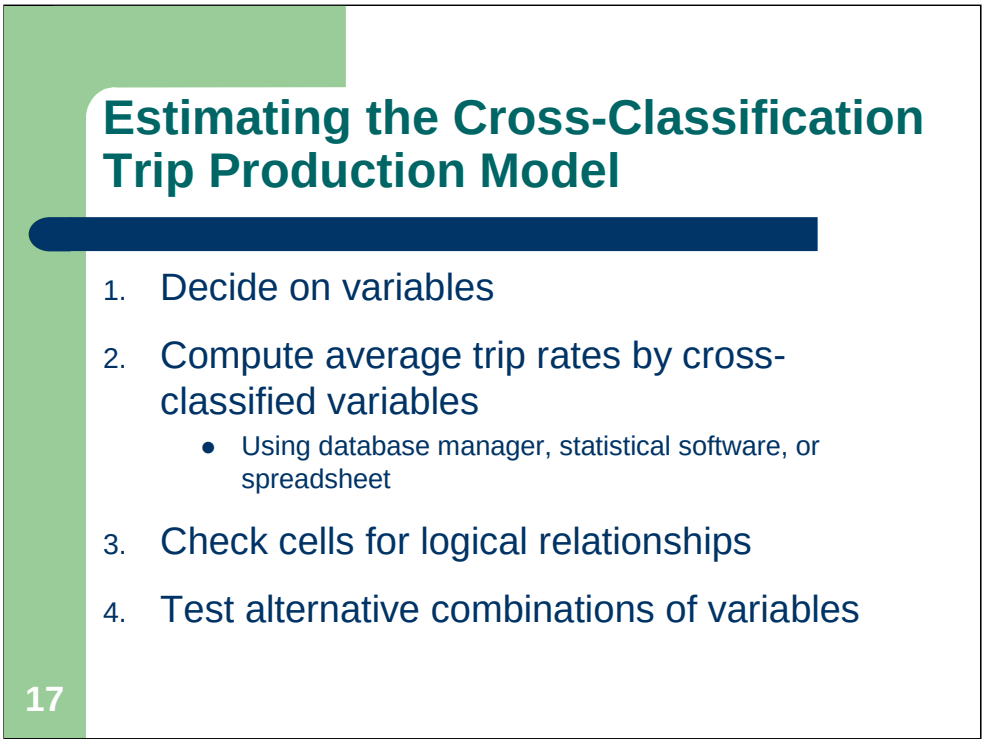
•Number of workers, which is used mostly for HBW and NHBW purposes

•Number of children, used for school and may be pickup/drop off trip purposes

•Number of autos, which is used for all purposes

•Income level, which once again can be used for all purposes.

All of these variables are generally available from most household surveys. Remember that forecasts of any variables chosen must be available for application. Note that typically for cross-classification models, we use two or more of the above variables. Most popular combinations include, Size * Autos, Size * Income Level, and Workers * Autos.



Key Contents: Estimating Cross-Class Rates

Details:

Next, we compute the average trip rates for each cell in the cross-classification table. Then, the estimated values must be checked for logical relationships. An example will be shown shortly.

It makes sense to test different combinations of variables to see which provides the best explanations for trip making behavior.

Example
Home Based Work Productions

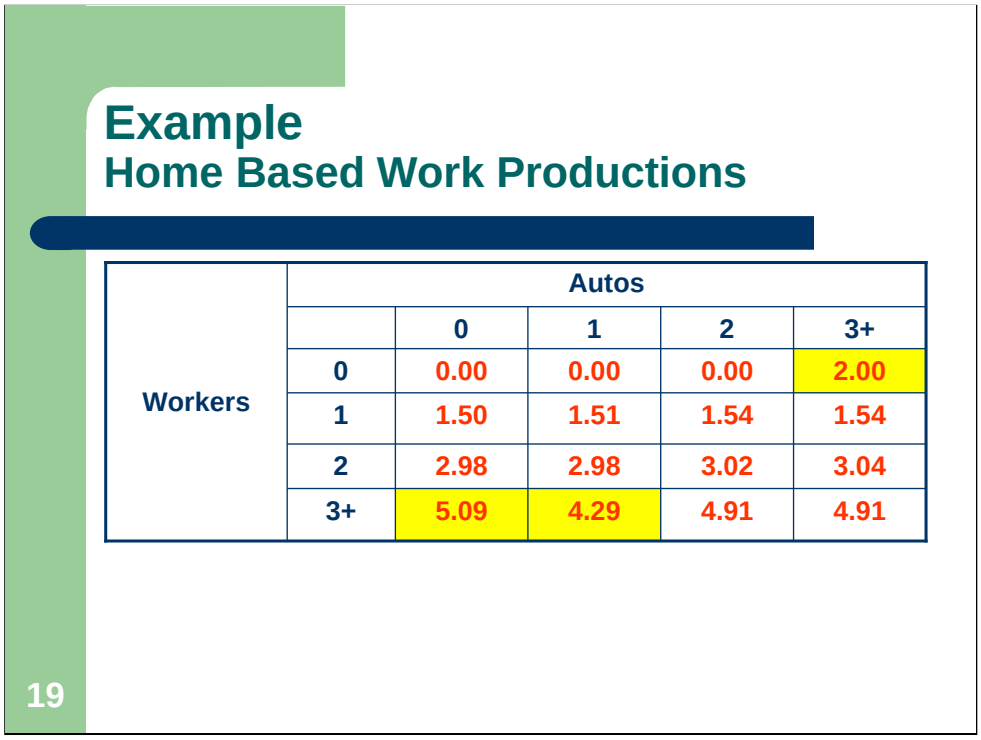
Workers	Autos			
	0	1	2	3+
0	0.00	0.00	0.00	2.00
1	1.50	1.51	1.54	1.54
2	2.98	2.98	3.02	3.04
3+	5.09	4.29	4.91	4.91

18

Key Contents: Estimating Cross-Class Rates – An Example

Details:

Here is an example of an initial estimation (before checking) of a home based work trip production model. Take a look for a moment and see whether you see any illogical relationships in the estimates.

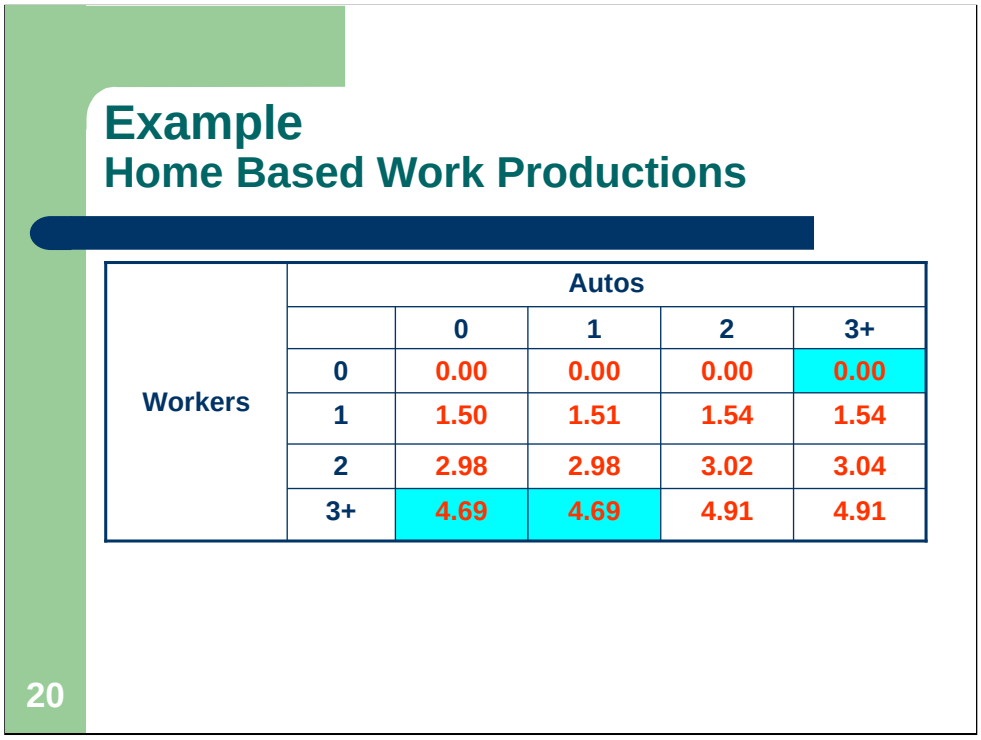


Key Contents: Estimating Cross-Class Rates – An Example

Details:

First, you may note the rate of 2.0 for 3+ person households with no workers. This seems illogical since while a non-worker may occasionally make a work trip (say, to a temporary position), it is unlikely that the rate would be as high as two trips per day. There is likely a low sample of households in that cell, and some error in reporting (trip purpose, worker status, etc.).

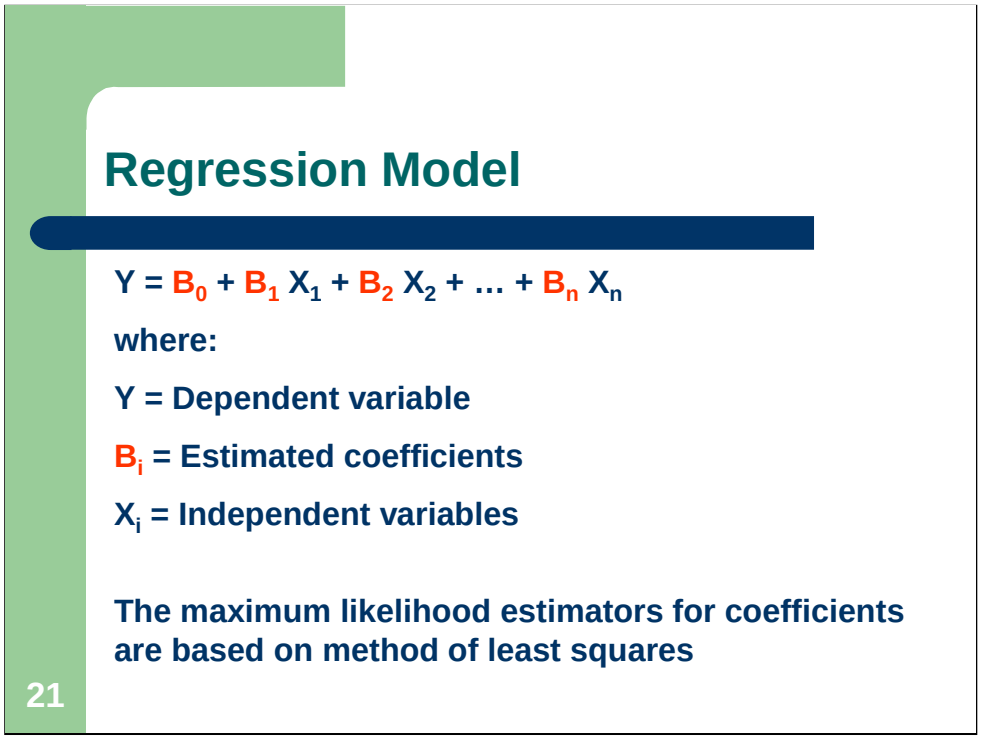
Also, it seems illogical that zero car households with three or more workers would make far fewer trips on average than one car households with 3+ workers. This is likely due to a small sample size for the 3+ worker/0 auto cell.



Key Contents: Estimating Cross-Class Rates – An Example

Details:

How do you handle these situations? In the case of the erroneous value, it can be replaced by a logical one. In the case of the small sample size causing issues with the trip rates, cells can be combined. In this case, the cells for 0 and 1 autos for 3+ workers were combined.



Key Contents: Regression Model

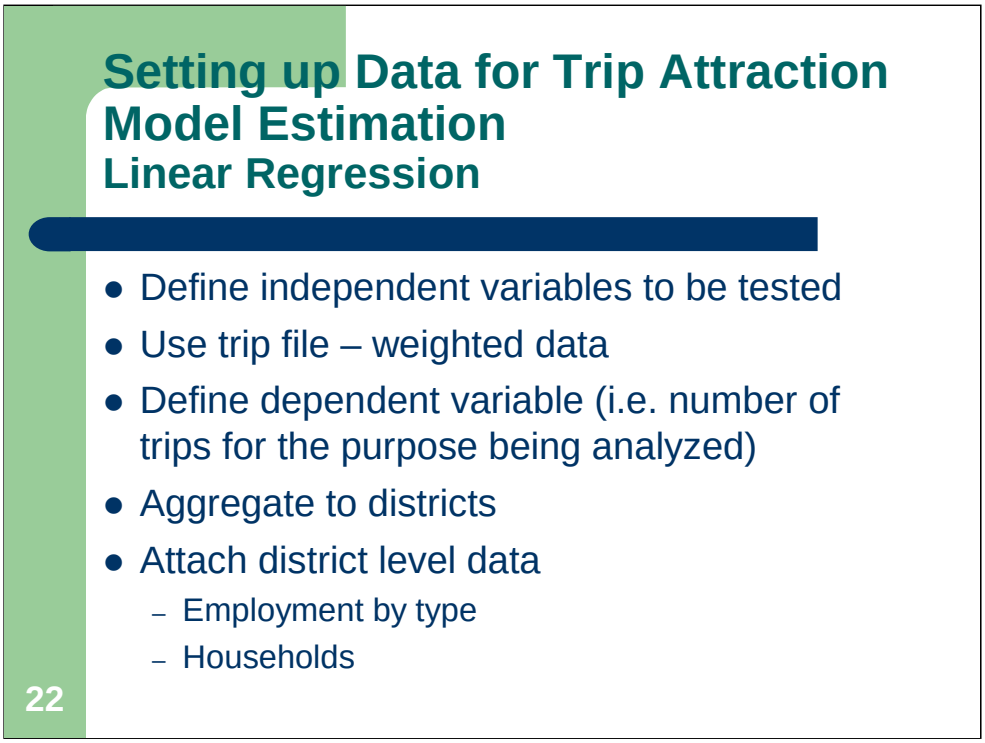
Details:

Now let's look at the regression model, which is typically used for trip attractions but can have other uses in travel models as well. This type of model was the basis for the homework from the last session.

Usually the number of trip attractions is a linear function of one or more variables describing the level of activity in a zone. Typically these values (the X_i's in the equation) are employment by type or the number of households. The objective of the estimation process is to estimate the values of the parameters (B_i).

It is logical to assert the intercept (B_0) to have a value of zero since it is illogical for a zone with no activity to attract any trips.

In linear regression, the least squares method produces the maximum likelihood estimates of the parameters.



Key Contents: Regression Model – Data Preparation

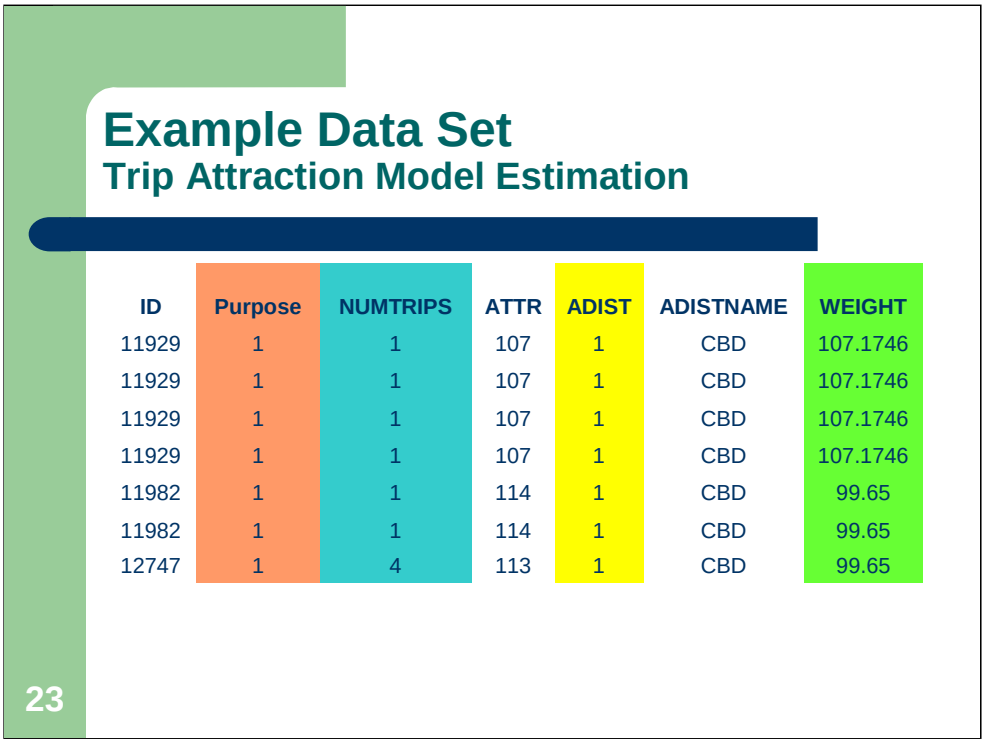
Details:

The first step in setting up the trip attraction model data set is to define the variables to be tested. It is not necessary to screen out potential problem variables initially; the estimation process will help determine the most significant and logical variables.

The household survey trip file with weights will be used. The dependent variable is the number of attractions for the trip purpose.

As was done in the homework exercise, the data must be aggregated to districts. This is because the typical sample sizes in household surveys are not sufficient to get good estimates of trips at the zone level. Districts should contain relatively similar types of zones and should be numerous enough (but not too numerous!) to come up with significant estimates for the coefficients.

The district level data set must be set up to include all potential variables for the model.

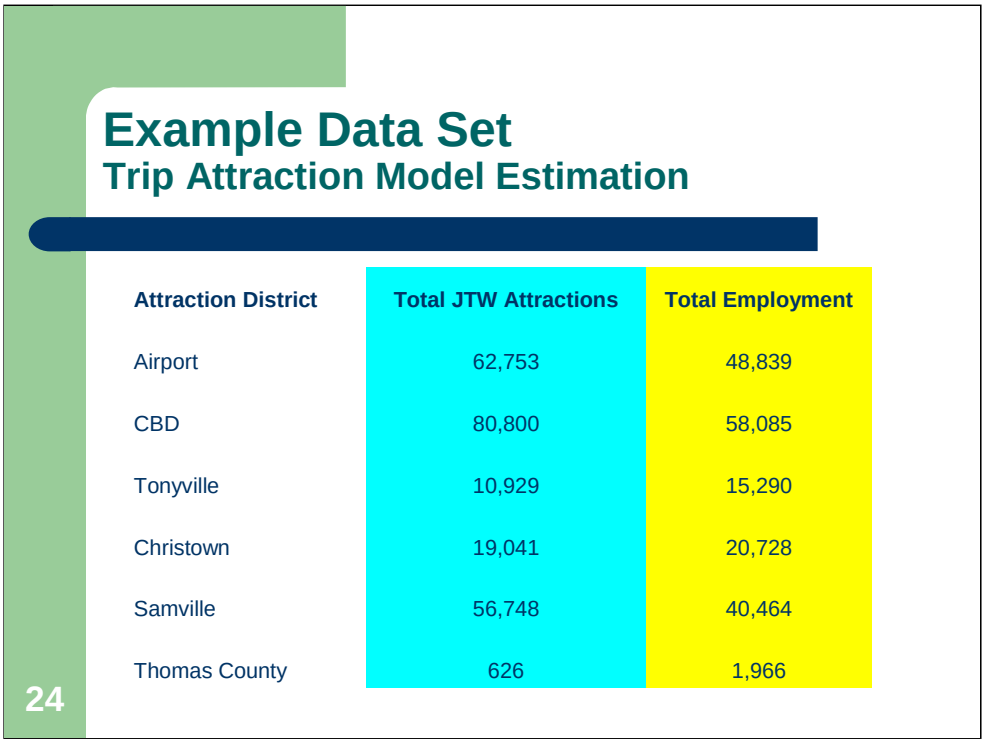


ID	Purpose	NUMTRIPS	ATTR	ADIST	ADISTNAME	WEIGHT
11929	1	1	107	1	CBD	107.1746
11929	1	1	107	1	CBD	107.1746
11929	1	1	107	1	CBD	107.1746
11929	1	1	107	1	CBD	107.1746
11982	1	1	114	1	CBD	99.65
11982	1	1	114	1	CBD	99.65
12747	1	4	113	1	CBD	99.65

Key Contents: Regression Model – Data Preparation Example

Details:

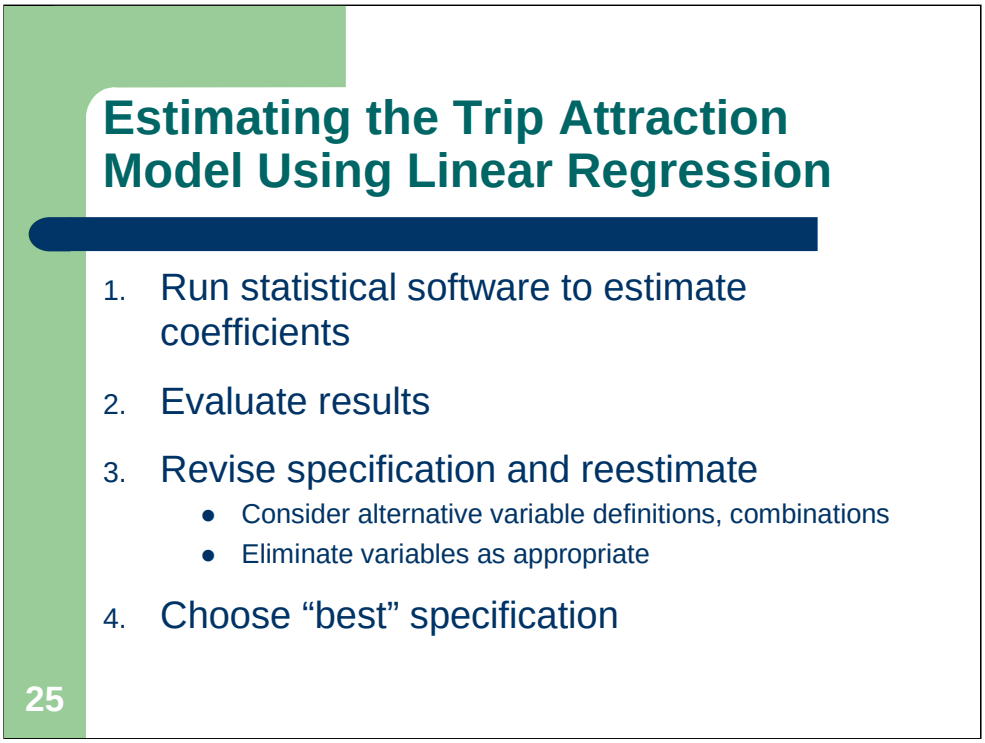
Here is an example of the original trip file with some of the important variables. ATTR represents the zone number for the attraction end of the trip; ADIST is the district number.



Key Contents: Regression Model – Data Preparation Example

Details:

This shows the data set aggregated to districts, ready for regression model estimation.



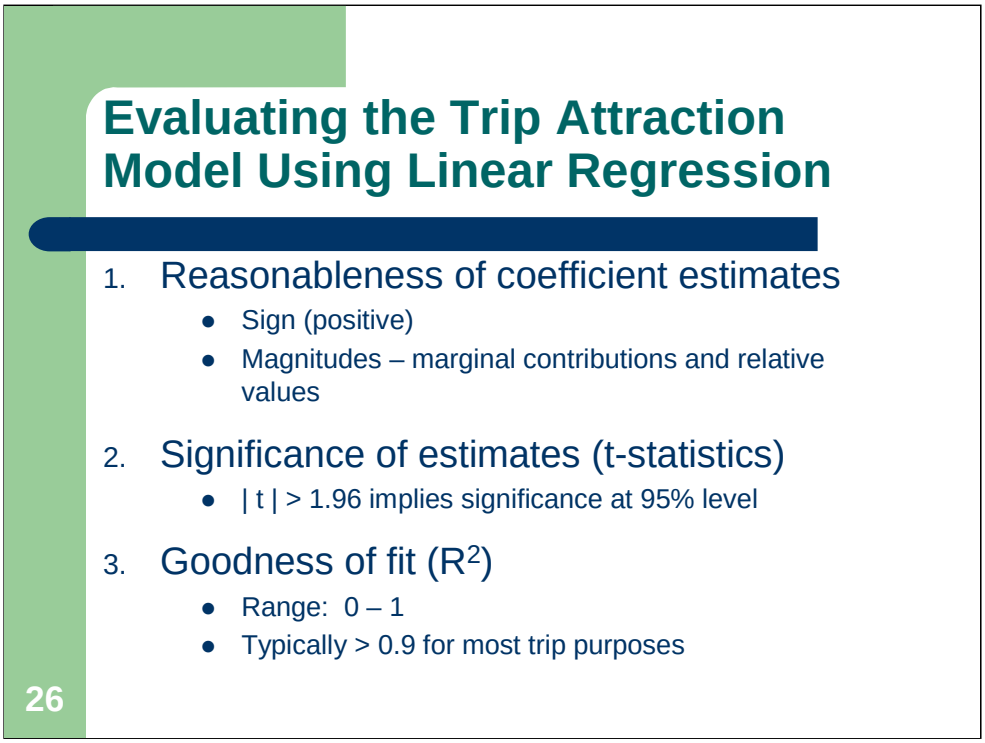
Key Contents: Regression Model – Parameter Estimation

Details:

It is relatively quick to run statistical software to obtain parameter estimates for the model. The results are evaluated, and changes made to the model to try to improve the results. These changes could include revising, combining, or eliminating variables. For example, if two variables were "service employment" and "government employment" and one of those variables' coefficient estimate had a reasonable while the other had an insignificant or negative value, a combined "service + government employment" variable could be tested, or the variable with the unreasonable coefficient could be dropped, depending on what makes sense.

It is also important to note that unreasonable results can be due to data problems. These results may indicate the need to recheck the data set.

When everything the modelers think of is exhausted, the best specification can be determined.



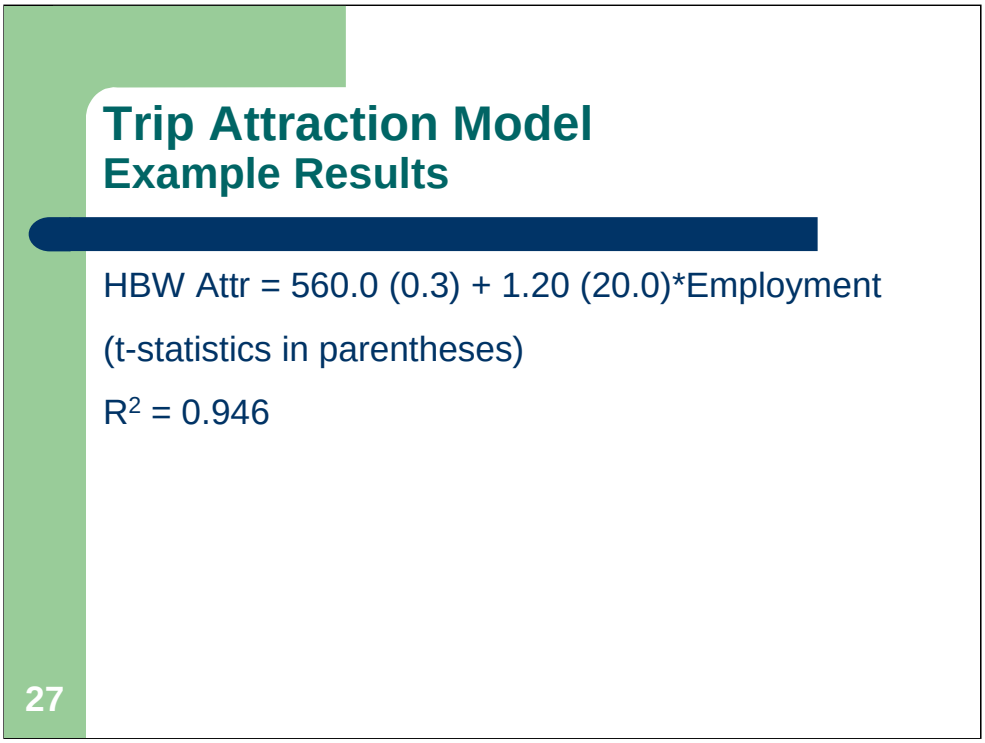
Key Contents: Regression Model – Reasonableness Checking

Details:

How do we check for reasonableness? In a trip attraction model, negative coefficients generally do not make sense. The marginal contributions for each variable should be considered. For example, does the number of trips per retail employee make sense relative to the number of trips per service employee?

Statistical significance of the coefficient estimates is tested using the t-statistics, as shown above. If coefficient estimates make sense, they are often retained even if the significance level is lower than 90%.

The R^2 value measures the goodness of fit for the entire model. While R^2 values of 0.9 or greater are typical for most trip purposes, lower values may be acceptable for purposes with relatively few trips.

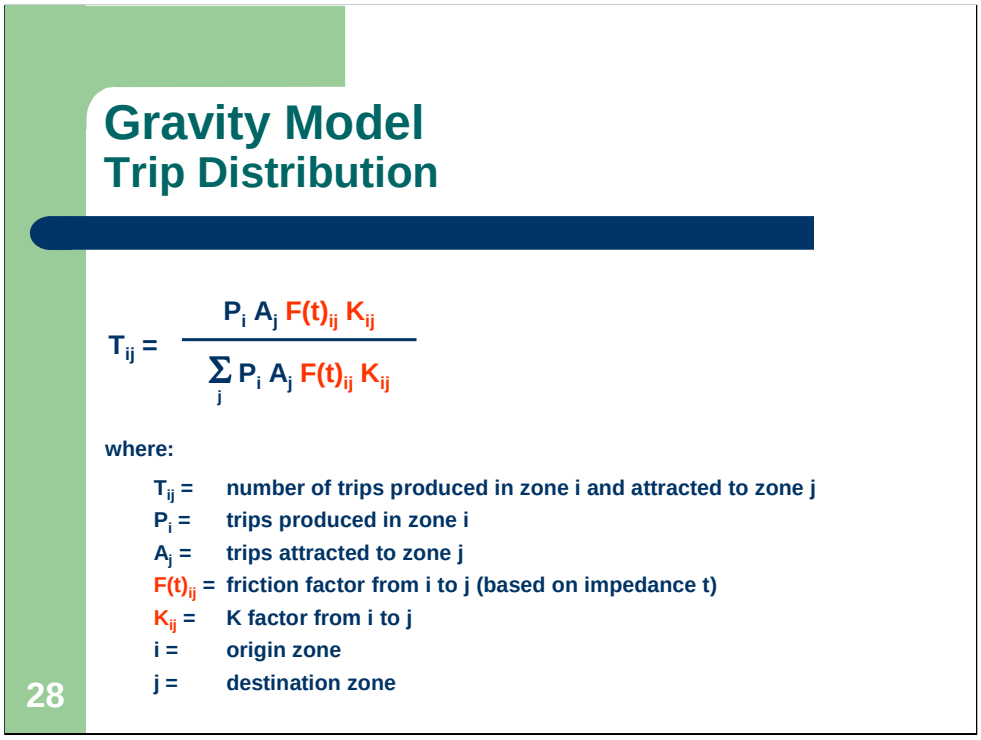


Key Contents: Regression Model – Reasonableness Checking Example

Details:

This example from the homework shows the following:

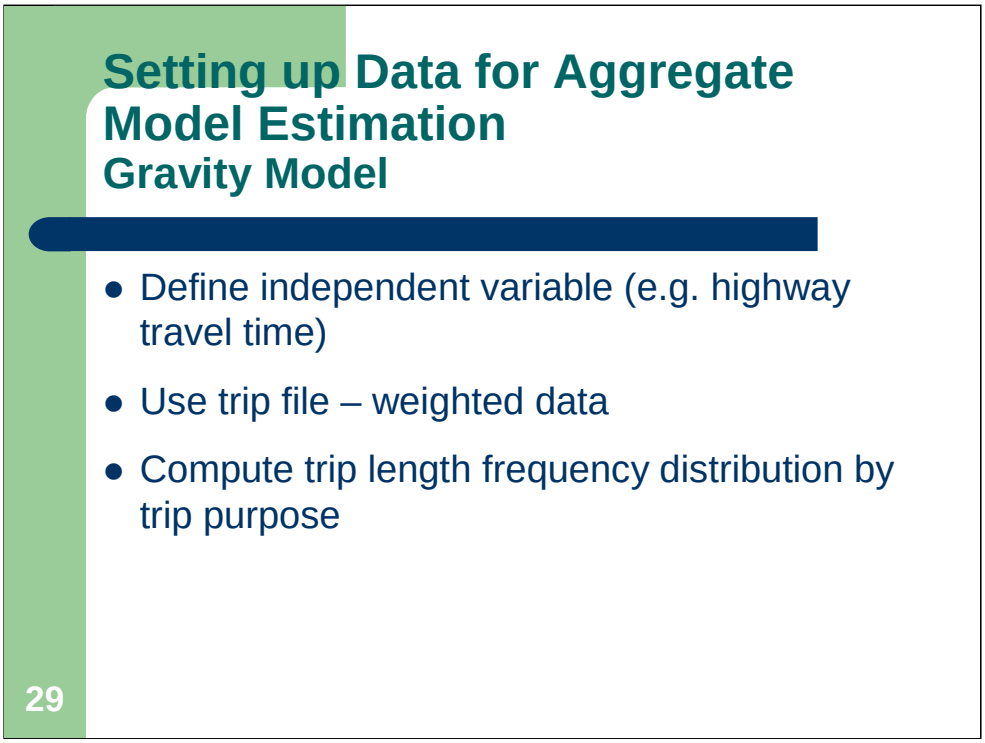
- The R^2 value is good, indicating a good fit.
- The t-statistic for the employment variable shows a very significant parameter (very unlikely for the true value to be zero).
- The estimate for the intercept is not significant at any reasonable level. It should be dropped from the model.



Key Contents: Gravity Models for trip Distribution

Details:

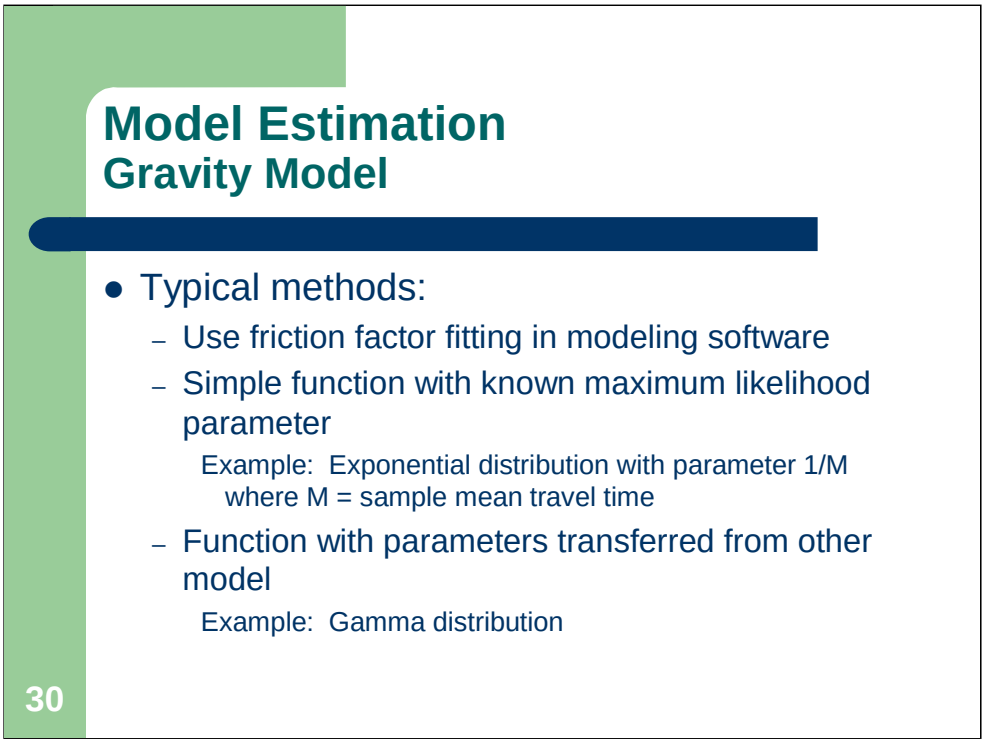
Next we consider the gravity trip distribution model. The estimated parameters are the friction factors, or the parameters of the distribution (e.g. gamma function) chosen for the friction factors. The K-factors are usually assumed to be 1.0 initially but may be adjusted during validation.



Key Contents: Gravity Models for Trip Distribution

Details:

The main components of setting up the data for the gravity model estimation are determining the impedance variable to be used and creating a trip length frequency distribution for that variable using the weighted trip data from the household survey.



Key Contents: Gravity Models – Typical Methods

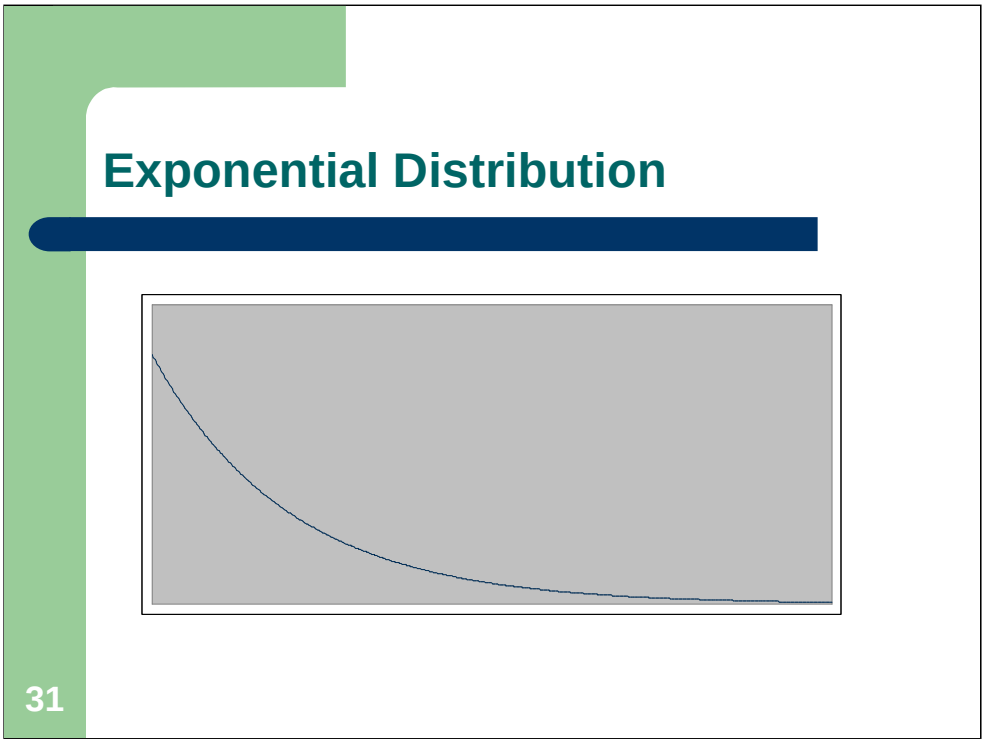
Details:

The typical methods for developing the gravity model parameters are:

• If the friction factors are simply fit to match the trip length frequency distribution (no function), the modeling software usually has procedures to estimated them.

• Some simple functions have known parameters that can be easily estimated from the data. For example, the exponential distribution with parameter 1/M where M = sample mean travel time.

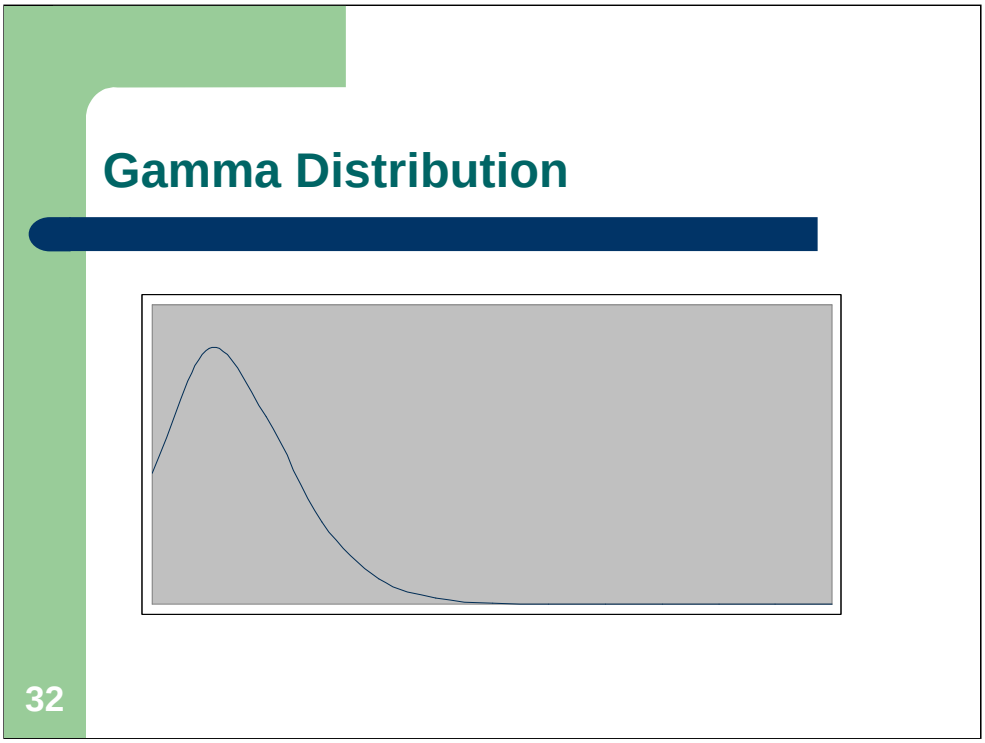
• If a more complex function is used, such as the gamma function, it is typical to transfer the parameters from another model.



Key Contents: Gravity Models – The Exponential Function

Details:

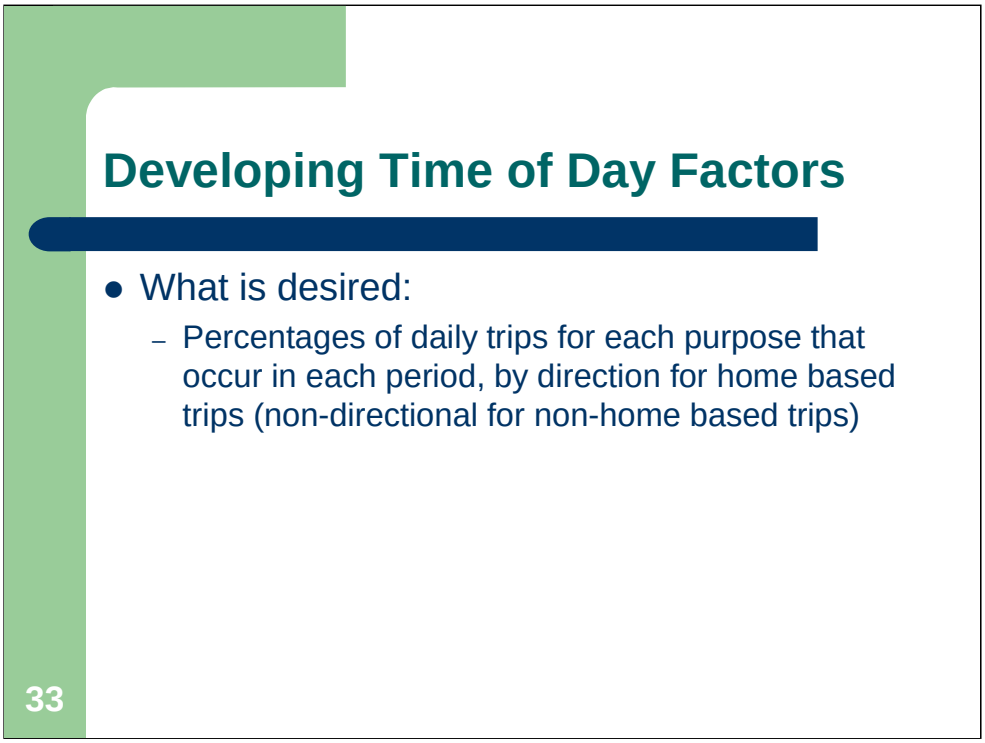
The exponential function is attractive because its parameter is estimated very simply and it has the property that lower factors are associated with higher travel times.



Key Contents: Gravity Models – The Gamma Function

Details:

The gamma function also has the property that lower factors are associated with higher travel times for most travel times. Very low travel times, which are associated with fewer trips, can have lower friction factors than for some higher travel times.



Key Contents: Time of Day Factors

Details:

Now let's consider the development of time of day factors. We seek to develop percentages of daily trips for each purpose that occur in each period, by direction for home based trips (non-directional for non-home based trips).

Setting up Data for Aggregate Model Estimation
Time of Day Model

- Determine resolution for testing (e.g. half hours)
- Use trip file – weighted data
- Define time variable (e.g. departure time, arrival time, midpoint)
- Define time periods

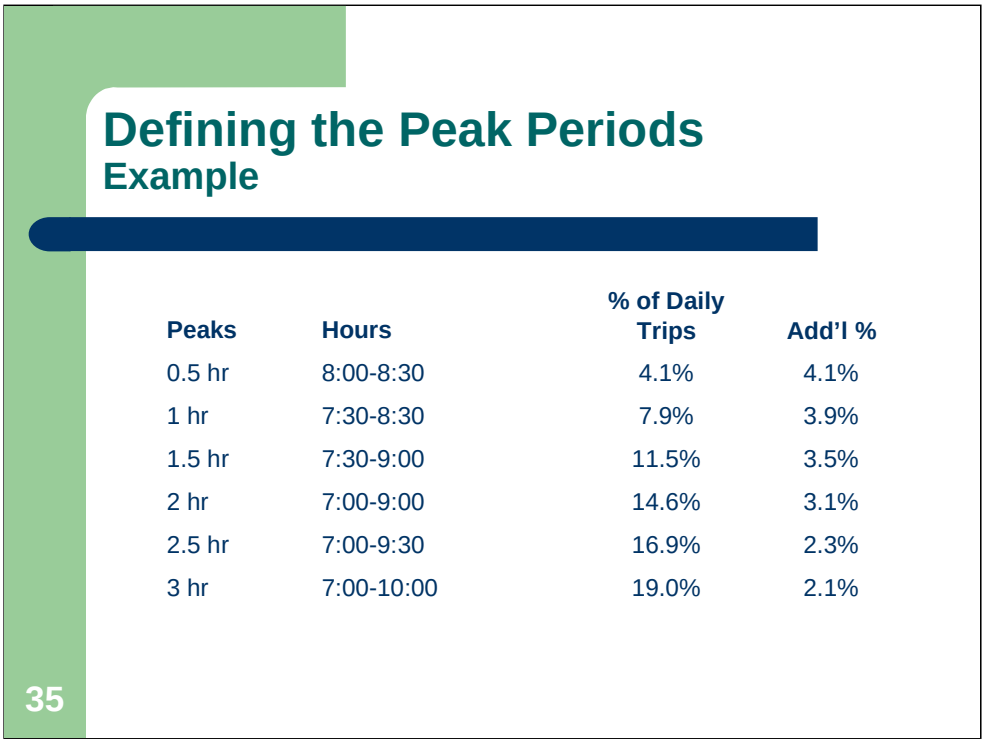
34

Key Contents: Time of Day Factors Estimation

Details:

First, we must define the time periods to be used. For example, there may be three periods, the a.m. peak, p.m. peak, and off-peak. These time periods can be defined from the weighted survey data, or alternate data sources (e.g. traffic counts) can be used, or a combination of sources can be considered. If the survey data are used, the data is divided into elemental periods (such as half hours or hours), and the percentages of trips in each period are compared to one another. The periods are subjectively defined from these data.

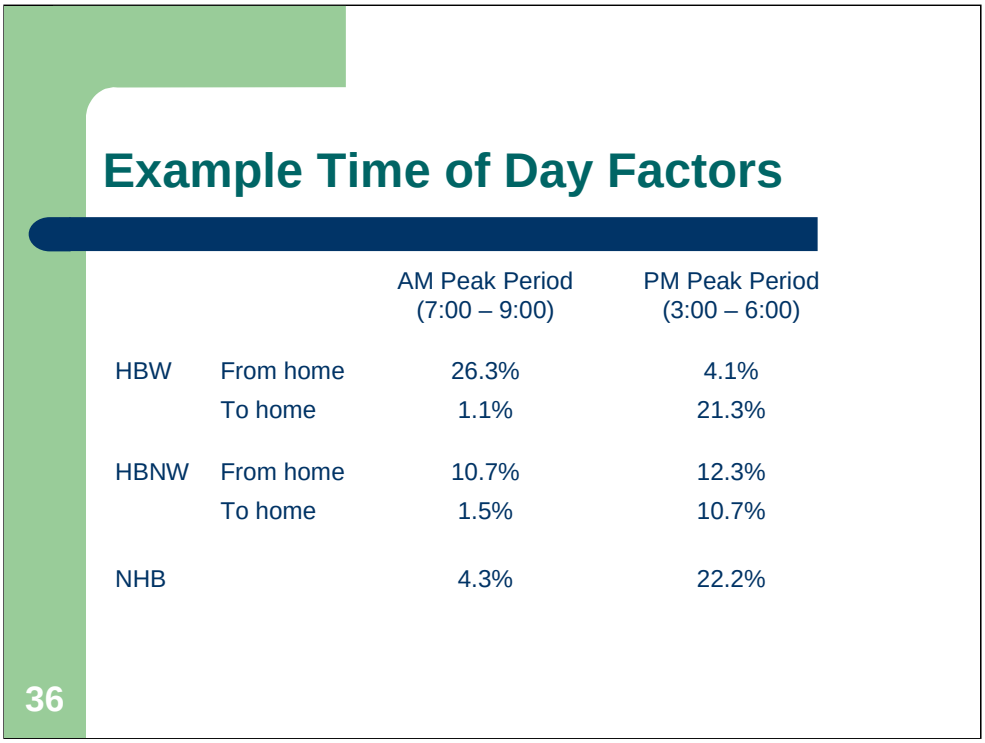
Some definition of which period a trip falls into must be made for trips that span two periods. The midpoint of the trip is a logical way to do this although the results do not vary much depending on this definition.



Key Contents: Defining Peak Periods

Details:

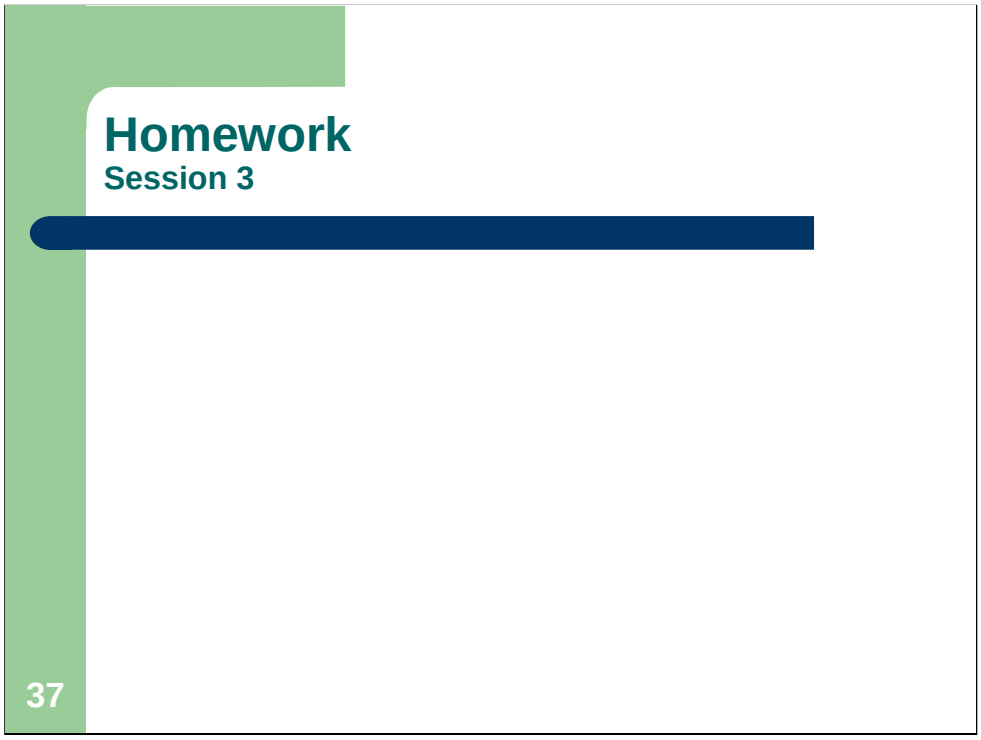
This is an example of how a time period may be defined. In this case, while the level of demand decreases with longer periods, it could make sense to define a two-hour peak period. Note that if a three-hour period were defined, the level of demand in the lowest half hour in the period is only half of that in the highest half hour.



Key Contents: Time of Day Factors - Example

Details:

Once everything has been defined, determining the percentages can be done easily suing a spreadsheet, database, or statistical software package. These are usually checked against other data sources such as traffic count data.



Key Contents: Homework 3

Details:

Please refer to the homework document for necessary instructions:

http://tmip.fhwa.dot.gov/discussions/webinars/archive/tmw/downloads/homework_3.pdf